Comparing Multiple Imputation and Propensity-Score Weighting in Unit-Nonresponse Adjustments: A Simulation Study

Alanya, Ahu; Wolf, Christof; Sotto, Cristina

This is a pre-copyedited, author-produced PDF of an article accepted for publication in
Public Opinion Quarterly, Vol. 79, 2015, No. 3, pp. 635-661
following peer review.
The version of record is available online at: https://doi.org/10.1093/poq/nfv029
Page numbers have been adjusted to the publishers version, whereby this postprint is fully quotable.

COMPARING MULTIPLE IMPUTATION AND PROPENSITY-SCORE WEIGHTING IN UNIT-NONRESPONSE ADJUSTMENTS

A SIMULATION STUDY

AHU ALANYA*
CHRISTOF WOLF
CRISTINA SOTTO

Abstract The usual approach to unit-nonresponse bias detection and adjustment in social surveys has been post-stratification weights, or more recently, propensity-score weighting (PSW) based on auxiliary information. There exists a third approach, which is far less popular: using multiple imputed values for each missing unit of the survey outcome(s). We suggest multiple imputation (MI) as an alternative to PSW since the latter is known to increase variance substantially without reducing bias when auxiliary variables are not associated with the survey outcome of interest. Given that most social surveys have multiple target variables, creating imputed data sets may address bias in survey outcomes with less variance inflation. We examine the performance of PSW and MI on mean estimates under various conditions using fully simulated data. To evaluate the performance of the methods, we report average bias, root mean squared error, and percent coverage of 95 percent confidence intervals. MI performs better under some of our scenarios, but PSW performs better under others. Even within certain scenarios, PSW performs better on coverage or root mean squared error while MI performs better on the other criteria. Therefore, robust methods that simultaneously model both the outcomes and the (non)response may be a promising alternative in the future.

Ahu Alanya is a doctoral candidate in the Centre for Sociological Research, University of Leuven, Leuven, Belgium. Christof Wolf is scientific director at GESIS Leibniz-Institute for the Social Sciences and professor of sociology at the University of Mannheim, Mannheim, Germany. Cristina Sotto is a professor at the Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Center for Statistics, University of Hasselt, Hasselt, Belgium. The authors thank the reviewers for their detailed and helpful comments. This work was supported by the FWO-Flanders [G0A0312N to A.A.]. *Address correspondence to Ahu Alanya, University of Leuven, Centre for Sociological Research, Parkstraat 45 - 3000 Leuven, Belgium; e-mail: ahu.alanya@ soc.kuleuven.be.

Until now, various indicators (for a summary, see Wagner [2012]) and adjustment methods (e.g., Groves 2006; Stoop et al. 2010; Bethlehem, Cobben, and Schouten 2011) have been suggested to detect and reduce unit-nonresponse bias in sample surveys. Among these approaches, propensity-score weighting (PSW) has become commonplace in survey research. Generally, this process computes propensity scores by modeling the probability/likelihood of the response indicator conditional on auxiliary information (e.g., sample frame information, paradata, or nonresponse surveys), then assigning each respondent a weight that is equal to the inverse of his/her estimated propensity score. Alternatively, respondents are classified into k equal-size strata based on estimated propensity scores, and a single nonresponse weight is computed for each stratum. Either way, PSW has the potential for high variance inflation and difficulty in handling item missingness in auxiliary variables effectively. However, it is a popular model-based technique for adjusting for unit nonresponse in survey research. Another missing-data method, multiple imputation (MI), has become one of the most attractive tools for item nonresponse adjustment (Rubin 1987). A number of previous simulation studies, particularly in the context of item nonresponse, have compared weighting adjustments to multiple imputation or robust extensions of the two (e.g., Kang and Schafer 2007; Cao, Tsiatis, and Davidian 2009), but only a few studies have actually applied MI to unit nonresponse. Simulations by Yuan and Little (2007) focused specifically on cluster samples and nonresponse mechanisms. Peytchev's (2012) study, on the other hand, provided practical evidence that multiple imputation for unit nonresponse can be more efficient compared to PSW in terms of standard errors. However, Peytchev's findings are based on a case study of a real-life survey with a relatively high response rate (above 70 percent), where he compares MI with only one kind of propensity-score-based method, namely, inverse propensity-score weighting. Thus, more exploration is needed on the relative efficiency of MI and PSW in unit-nonresponse adjustment under various adjustment scenarios, using commonly available versions of MI and PSW methods.

MI may offer several advantages over PSW. One advantage is related to efficiency. PSW tends to inflate variance estimates, particularly when auxiliary variables used in the adjustment are strongly associated with the response propensity and not associated with the outcome (Little 1986; Little and Vartivarian 2005). Although outcome-specific propensity-score models could produce some efficiency gains by excluding auxiliaries not related to the survey outcome of interest, this is not feasible for general social surveys, which target a broad range of survey variables. Similarly, trimming inverse propensity-score weights (Lee, Lessler, and Stuart 2011) or using stratification on propensity scores (Rosenbaum and Rubin 1984) can help reduce variance by avoiding large weights, but also can increase bias if employed rigorously. Multiple imputation, on the other hand, can be equally good at addressing bias

and yielding lower standard errors even when there are a number of auxiliary variables and adjustment targets multiple survey outcomes. Furthermore, the new methods developed under MI allow for separate modeling of the response propensity and the missingness in the outcome of interest; consequently, auxiliary variables related to either response or outcome or both are combined properly under MI (Jolani, van Buuren, and Frank 2011).

It is common to have missing values in auxiliary variables in unit-nonresponse adjustment, and this should be resolved before estimating propensity scores. An effective way to solve this problem is to use multiple imputation as an initial step before weighting to complete missing auxiliary information (e.g., Mattei 2009; Qu and Lipkovich 2009). However, this method needs further examination, particularly on the correct estimation of standard errors after using MI as an initial step and on modeling propensity scores across the imputed data sets. Other ways to estimate propensity scores with missing data, such as those offered by D'Agostino and Rubin (2000), also require some extra work and are not available in mainstream software. Multiple imputation, on the other hand, can handle item missingness in auxiliary variables and unit nonresponse in one step. In addition, with increasing administrative data, paradata, and other auxiliary information, the missingness patterns in unit-nonresponse adjustments are becoming more complex (less monotone). Therefore, the ability to simultaneously account for item and unit nonresponse may become a more apparent advantage for MI over PSW (Little 2013).

In this paper, we explore whether MI is a better alternative to PSW in unit-nonresponse bias adjustment in terms of providing lower root mean square errors (RMSE) and/or higher coverage of 95 percent confidence intervals. For this purpose, we investigate how MI and PSW estimates differ in relation to the amount of missing data, as well as the strength of associations between auxiliary variables, the response propensity, and outcome variables. Furthermore, we consider how robust each method is against misspecifications such as omitted interactions and nonlinear terms, which can be especially meaningful in survey research where “main-effects” weighting models appear to be a common practice. Moreover, our comparison between MI and PSW considers two different versions of each method. For example, subclassification on propensity scores may yield variance estimates closer to that of MI compared to inverse propensity-score weighting. This could make switching to MI superfluous given the current expertise in propensity-score-based methods. Finally, we restrict attention to nonresponse of a missing-at-random (MAR) nature, as explained in the following section. It is worth noting that this study focuses on model-based approaches to unit nonresponse, leaving out traditional techniques such as complete case analysis, post-stratification weighting, or relatively new robust techniques that are not currently incorporated into mainstream software packages. That is, we use off-the-shelf Stata routines for PSW and MI that are widely available to analysts and data providers alike. As Little (1988, 288) suggests, “carefully

constructed" nonresponse adjustments with model-based methods can improve our analysis compared to traditional approaches; however, practitioners should be aware of the benefits as well as the risks associated with alternative methods. This study specifically addresses the modeling questions that practitioners ask themselves when using either of these model-based approaches; it aims to provide insight on bias-variance trade-offs and the resulting coverage for population parameters under different conditions.
Using simulated data, we compare the performance of MI and PSW under varying levels of the factors discussed above (response rate, strength of associations, adjustment model misspecification) to answer the following research questions:

1. Can multiple imputation (MI) yield consistently lower variance estimates compared to propensity-score weighting (PSW) while being equally effective in reducing bias?
2. How do the bias-variance properties of MI and PSW affect 95 percent confidence interval coverage?
3. How do the methods perform under different response rates, different degrees of associations with auxiliary variables, different sample sizes, and different specifications of the model?

Simulation Study

SIMULATION SETUP

A number of simulation studies have investigated the properties of PSW methods (e.g., Brookhart et al. 2006; Kreuter and Olson 2011). In addition, some other studies have compared the performance of inverse propensity-score weighting (IPSW) to MI for item nonresponse. For example, Carpenter, Kenward, and Vansteelandt (2006) assessed, for continuous outcomes, a doubly robust IPSW estimator with standard IPSW, maximum likelihood, and MI. IPSW estimators were found to be inefficient and sensitive to the choice of the weight model, but the doubly robust version was as efficient as MI and robust against misspecification. Beunckens, Sotto, and Molenberghs (2008), on the other hand, considered IPSW and MI-based approaches for binary longitudinal outcomes. Their simulations underscored the sensitivity of IPSW to misspecification in the weight model and its inefficiency for modest amounts of missingness. Moreover, in all scenarios investigated, the MI-based approach outperformed the weighted approach despite misspecification in either the imputation model or the analysis model.
Building on these studies, our simulations compare PSW and MI in the case of unit nonresponse. We focus on the estimation of the population mean of a continuous variable from incomplete data. Accordingly, hypothetical survey data sets are generated (with sample sizes n = 200, 3,000, and 10,000), where

the survey outcome of interest is Y and the binary unit-response indicator is R. For simplicity, we consider two auxiliary variables, Z_1 and Z_2, independently drawn from the standard normal distribution ~ N (0, 1). R is then modeled as a function of Z_1 and Z_2, either as

$$\Pr\left(\mathrm{R}=1|\mathrm{Z_i}\right)=\frac{e^{\mathrm{a_0}+\mathrm{a_1}Z_1+\mathrm{a_2}Z_2}}{1+e^{\mathrm{a_0}+\mathrm{a_1}Z_1+\mathrm{a_2}Z_2}} \tag{1a}$$

or

$$\Pr\left(\mathrm{R}=1|\mathrm{Z_i}=\frac{e^{\mathrm{a_0}+\mathrm{a_1}Z_1+\mathrm{a_2}Z_2+\mathrm{a_3}Z_1^2+\mathrm{a_4}Z_1Z_2+\mathrm{a_5}Z_2^2}}{1+e^{\mathrm{a_0}+\mathrm{a_1}Z_1+\mathrm{a_2}Z_2+\mathrm{a_3}Z_1^2+\mathrm{a_4}Z_1Z_2+\mathrm{a_5}Z_2^2}}\right. \tag{1b}$$

respectively reflecting a “main-effects” or a “complex” formulation for the underlying association between R and Zs. In the next step, a continuous survey outcome Y is modeled as a function of Z_1, Z_2, a constant term, and a normally distributed error term e ~ N(0,1). As in the case of equations (1a) and (1b), we formulate either a “main-effects” or a “complex” model as well, i.e.,

$$\mathrm{Y}=10+\beta_1Z_1+\beta_2Z_2+\beta_e\mathrm{e} \tag{2a}$$

or

$$\mathrm{Y}=10+\beta_1Z_1+\beta_2Z_2+\beta_3Z_1^2+\beta_4Z_1Z_2+\beta_5Z_2^2+\beta_e\mathrm{e} \tag{2b}$$

To make our scenarios more realistic, we manipulate the variance explained in Y by varying the coefficient of the error term ¡5e in equations (2a) and (2b). These equations lead to, in total, two different synthetic data sets where the first models Z, R, and Y relationships using only main effects and the second adds quadratic and interaction terms. We do not include R as a predictor in the data-generation model of the outcome variable Y. That is, we assume that R has no direct effect on Y (as shown as a dashed arrow in figure 1) independent from the auxiliary variables, implying that the nonresponse mechanism is MAR (Rubin 1976). Regarding other design features, we use a similar

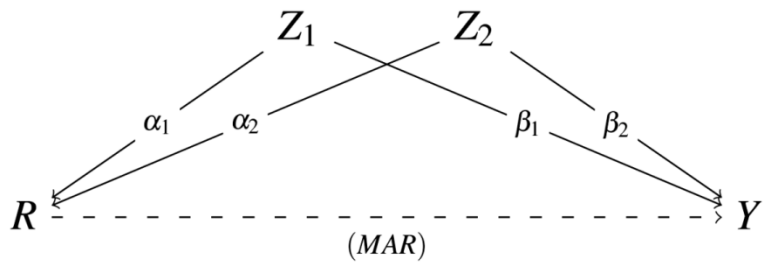

Figure 1. Data-Generation Model – Illustrated for the Main Effects Scenario.

simulation framework introduced in previous studies particularly by Setoguchi and his colleagues (2008) in the area of epidemiological research, and Kreuter and Olson (2011) as applied to survey methodology. We set the average of the survey variable Y to 10 and vary the strength of auxiliary-response and auxiliary-outcome relations. To create weak, moderate, and strong associations between the auxiliary variables and the response propensity, we consider possible combinations of {0.1,2,4} for α_1 and α_2 in equations (1a) and (1b), whereas α_0 is used to control the overall amount of missing data. Similarly, in equations (2a) and (2b), we use combinations of {0.1,1,3} for $\beta_1$and β_2 to represent different degrees of association between the auxiliary variables and the outcome Y.
For conciseness, we summarize our findings by discussing the results for only three patterns (or combinations of αs and βs).[1] We organize our simulations and results primarily by the patterns of association between R, Y, and auxiliary variables. In the first pattern, auxiliary variables are weakly associated with both R and Y ($\alpha_1 = 0.1, \alpha_2 = 0.1, \beta_1 = 0.1, \beta_2 = 0.1$).Therefore, bias and variance inflation as well as the difference between the methods are expected to be small (see Kreuter et al. [2010]). For nonresponse bias adjustment to be effective, auxiliary variables need to have substantive association with both R and Y. However, most auxiliary data collected at the time of the survey target understanding the response behavior and therefore are likely to have strong correlations with R rather than with the Ys. Furthermore, general practice with PSW is to use variables that significantly predict R, since weighting is applied for multiple outcomes and includes auxiliary variables that have different levels of associations with the Y variables (e.g., see Matsuo et al. [2010]). In this case, the practitioner may end up with the second pattern, where auxiliary variables are strongly related with R, but weakly determine Y ($\alpha_1 = 2, \alpha_2 = 4, \beta_1 = 0.1, \beta_2 = 0.1$) ,for which we expect PSW to inflate variance more than MI methods. The last pattern is the ideal adjustment case and is one where auxiliary variables have high levels of associations with both the response and the outcome ($\alpha_1 = 0.1, \alpha_2 = 0.1, \beta_1 = 1, \beta_2 = 3$). We expect higher bias and variance as we go from pattern 2 to pattern 3 since the adjustment gets stronger. Given that situations where auxiliary variables are strongly related to Y but weakly related to R are less likely to occur in surveys with multiple outcomes, we exclude this pattern in our simulations. Full details on data generation and the three patterns of coefficients are presented in appendix 1. Overall, we expect PSW and MI to perform similarly in correcting bias, with the former having higher variance inflation compared to the latter, particularly for patterns 2 and 3.
To investigate how differences between PSW and MI are affected by the amount of missing data, we also vary the response rate. Response rates of

[1] These three patterns of associations between response (R), outcome (Y), and auxiliary variables are crucial in the comparison of MI and PSW methods. In addition, these patterns are easy to examine before researchers decide on a certain method.

face-to-face general-population social surveys today can be as low as 30 percent (e.g., 34 percent for ALLBUS 2010, 31 percent for ESS 2010 in Germany, 29 percent in ISSP 2009 in Argentina, and so on). When the response rate is low, the rates of missing information and bias are expected to be larger and these may have an impact on how the methods perform. For example, strong auxiliary variables are likely to result in larger weights under lower response rates, yielding larger variance inflation for PSW. Also, if the rate of missing information for the parameter of interest is high, more than five imputed data sets may be necessary to achieve efficiency (Schafer 1999). Therefore, we generate two different levels of response rates that we think are relevant for survey practitioners: low ($\simeq$ 35 percent) and moderate ($\simeq$ 65 percent) by manipulating the constant term a0 in the true response model (see appendix 2 for the resulting response rates).

Simulation Scenarios

Simulation studies have focused on either the direction/strength of the relationship among auxiliary variables, response propensity, and outcome variables (e.g., Brookhart et al. 2006; Kreuter and Olsen 2011) or the (mis)specification of propensity-score models (Drake 1993; Millimet and Tchernis 2009; Clarke, Kenkel, and Rueda 2011); our paper focuses on both the strength of the relationship of auxiliary variables with R and Y as well as the (mis)specification of propensity scores. We have defined three different patterns between auxiliaries and R and Y to vary these associations. To vary the functional form of the adjustment, we created realistic complexity in the true propensity-score function by adding nonadditive and nonlinear terms in the true response and outcome models. Also, because focusing on the main effects of auxiliary variables that are significant predictors of the response indicator when specifying propensity scores often results in misspecification of the possibly more complex true response models, we consider the effect of misspecification of adjustment models for each synthetic data set, as shown in table 1.
Overall, we consider four adjustment scenarios:

Scenario 1.1: A main-effects model, in which the functional form of the true response model includes only main effects and adjustment models are correctly specified;

Scenario 1.2: A misspecified main-effects model, where the functional form of the true response model includes only main effects and adjustment models are misspecified by omitting one of the auxiliary variables;

Scenario 2.1: A complex model, where the functional form of the true R model is complex, and includes quadratic and interaction terms, and adjustment models are correctly specified;

Table 1. Simulation Scenarios

Scenario	
	True response model
1.1,1.2	logit (Pr(R = 1)) = $\alpha_0 + \alpha_1 Z_1 + \alpha_1 Z_2$
2.1,2.2	logit (Pr(R = 1)) = $\alpha_0 + \alpha_1 Z_1 + \alpha_2 Z_2 + \alpha_3 Z_1^2 + \alpha_4 Z_1 Z_2 + \alpha_5 Z_2^2$
	True survey outcome model
1.1,1.2	$Y = 10 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_e^e$
2.1,2.2	$Y = 10 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_1^2 + \beta_4 Z_1 Z_2 + \beta_5 Z_2^2 + \beta_e^e$
	PSW (π) model
1.1	logit(R)= f(Z_1,Z_2)
1.2	logit(R)= f(Z_1)
2.1	logit(R)= f(Z_1,Z_1,Z_1^2,Z_1,Z_2,Z_2^2)
2.2	logit(R)= f(Z_1,Z_2)
	MI (regression) model
1.1	mi(Y)= f(Z_1,Z_2)
1.2	mi(Y)= f(Z_1)
2.1	mi(Y)= f (Z_1,Z_2,Z_1^2,Z_1.Z_2,Z_2^2)
2.2	mi(Y)= f(Z_1,Z_2)

Scenario 2.2: A misspecified complex model, in which the functional form of the true R model is complex and adjustment models are misspecified by omitting quadratic and interaction terms. Table 1 summarizes the true underlying models (data-generation models) and adjustment models for each scenario.

PSW AND MI METHODS

Stata 12 software was used for all tasks described. First, we generated the data and saved the complete data estimates for the mean and variance of Y. To generate incomplete data sets, Stata deleted Y values for nonrespondents who had a value of 0 for the response indicator (R). Next, the program implemented four types of unit-nonresponse adjustment: two propensity-score-based methods and multiple imputation with two different numbers of imputed data sets equal to 5 and 100. A brief description of the methods along with their abbreviations is provided below.

IPSW, inverse propensity-score weighting, was applied using the inverse of the estimated propensity scores from a logistic regression of R on the auxiliary variables as weights for respondents (e.g., Hirano and Imbens 2001). Stata's user-written command pscore was used to estimate the response propensities. Inverse of response propensities (1/p) were included as (unnormalized) weights using the pweight command of Stata for estimating means and variances based

on the respondents' sample. Second, we generated weights using propensity-score subclassification (PSS) to minimize the variance impact of PSW (Little and Rubin 2002). We followed the same rules described by Little (1986) and also employed by Lee and Valliant (2009). The total sample of respondents and non-respondents was divided into ten equal-size groups based on the estimated propensity scores from IPSW. Weights for respondents were calculated as the ratio of sample units within each strata to the total number of units in the entire sample divided by the proportion of respondents in each stratum to the total number of respondents in the entire sample. For example, weights in the i^{th} stratum are expressed as

$$W_i^R = \frac{(n_i^{NR} + n_i^R) \,/\, (n^{NR} + n^R)}{(n_i^R / n^R)} \tag{3}$$

where i is the i^{th} stratum, n_i^R is the total number of respondents in the i^{th} stratum, n_i^{NR} is the total number of nonrespondents in the i^{th} stratum; and n^R is the total number of respondents and n^{NR} is the total number of nonrespondents in the entire sample.

The other two methods, MI with five imputed data sets and with 100 imputed data sets, were implemented using the chained equations approach (Raghunathan et al. 2001) as opposed to conventional MI. While the latter assumes a multivariate normal distribution (mvn) for the multivariate outcomes, the former is less stringent in the sense that only univariate normality of one outcome conditional on the other outcomes (in some specific order, i.e., “chained”) is required. When there are relatively few variables to impute and the variables to be imputed are approximately jointly normally distributed, it is convenient to use the mvn method. However, this is rarely the case in social surveys that may include a wide array of variables. The advantage of the chained equations approach is that it can handle various data characteristics, such as ranges or ordinal scales, using conditional models (White, Royston, and Wood 2011).[2] We use the chained equations method even though the only missing data are in Y, so the result is univariate regression.

Unlike weighting, MI models the target survey variable instead of the response propensity. Missing Y values are imputed m times by taking independent random draws from the posterior predictive distribution of the missing data (Y_{miss}) given the observed data (Y_{obs}). The number of required imputations to adjust for bias is typically as low as five, but since we also investigate low response rates that reflect high rates of missingness, we considered m = 100 imputations to evaluate relative efficiency gains from increasing the number of imputations.

[2] In principle, the chained equations approach is thus more suitable for social surveys, although there are studies suggesting that MI assuming mvn performs well even with non-normal variables (Lee and Carlin 2010). However, this advantage is not evident in our simulations since we consider only a single outcome (i.e., univariate) rather than multiple outcomes (i.e., multivariate).

PERFORMANCE METRICS

The simulation scenarios and patterns are designed to assess how methods perform in terms of bias-variance trade-offs under different conditions. We looked at two major performance metrics: bias and RMSE. For a given method, we calculated these measures as

$$\text{bias}\left(\bar{\gamma}_{\text{R}}\right) = \sum_{\text{i=1}}^{\text{S}} \frac{\bar{\gamma}_{\text{iR}}}{\text{S}} - \mu = \bar{\gamma}_{\text{R}} - \mu \tag{4}$$

$$\text{var}\left(\bar{\gamma}_{\text{R}}\right) = \sum_{\text{i=1}}^{\text{S}} \frac{\text{var}\left(\bar{\gamma}_{\text{iR}}\right)}{\text{S}}, \text{average of variances} \tag{5}$$

$$RMSE\ \left(\bar{\gamma}_{R}\right) = \sqrt{bias^2(\bar{\gamma}_{R})} + \qquad \left(\bar{\gamma}_{R}\right), \tag{6}$$

where S = 500 denotes the number of samples drawn from each population, $\bar{\gamma}_{\text{R}}$ is the adjusted or unadjusted respondent mean of Y, and $\bar{\gamma}_{\text{iR}}$ is the respondent mean for the i^{th} simulated data set, with variance estimate var ($\bar{\gamma}_{\text{iR}}$) as provided by the software; $\bar{\gamma}_{\text{R}}$ is the average of the adjusted or unadjusted respondent means across all simulated data sets. Bias is calculated for each simulation cell and averaged over S replications. Similarly, variance for each pattern and scenario is calculated by averaging variance estimates over S replications. Additionally, to evaluate whether the efficiency of MI results in lower coverage for the true population mean, we report the percent coverage of the 95 percent confidence intervals. The latter is calculated as the percentage of 500 simulated data sets in which the adjusted or unadjusted respondent sample 95 percent confidence interval includes the population mean.

Results

Tables 2-4 show the performance metrics for each pattern of association by scenario under low and moderate response rates and for n = 200, 3,000, and 10,000, respectively. One of the questions posed is whether MI approach is better than PSW in reducing bias while producing lower variance estimates. The answer from the tables is not straightforward. MI generally yields lower RMSE when auxiliary variables are strongly associated with both R and Y (pattern 3). However, this does not always result in higher coverage, for example when the true R and Y models are complex, but this complexity is ignored in the adjustment model (scenario 2.2). Overall, the tables show that MI is not consistently better than PSW. Therefore, it seems reasonable to present the results by pattern and specify where each method may have strengths and the implications of this for future practice and research. Below, we present a more detailed discussion with reference to smaller and larger data sets, taking the n = 3,000 as the main reference point.

Table 2. Performance Metrics by Pattern, Levels of Associations with Auxiliary Variables (500 simulated data sets of n =200)

		Low response rate											
		Pattern 1				Pattern 2				Pattern 3			
Metric	Scenario	IPSW	PSS	MI5	MI100	IPSW	PSS	MI5	MI100	IPSW	PSS	MI5	MI100
Bias $x10^3$													
	1.1	-7	-7	-17	-13	63	93	56	0	1302	1877	288	8
	1.2	1	0	-34	-4	103	105	68	94	2980	2981	2963	2964
	2.1	-6	-2	6	-9	3	52	105	-3	846	1586	531	-9
	2.2	24	15	14	16	187	74	-136	-179	292	1483	-573	-790
RMSE $x10^3$													
	1.1	127	133	131	125	213	199	288	227	2492	2338	1935	1632
	1.2	126	132	149	129	168	175	160	164	3166	3187	3014	3122
	2.1	131	138	151	135	220	189	582	448	2365	2168	3724	3139
	2.2	133	138	135	129	294	209	296	279	3720	2161	1962	1820
95% CI cov.													
	1.1	93	92	93	93	81	87	93	94	51	44	92	94
	1.2	93	94	91	93	87	86	88	88	2	3	8	3
	2.1	95	95	94	95	83	91	97	96	65	53	97	96
	2.2	93	94	91	93	51	89	87	84	43	58	91	87

Continued

Table 2. Continued

		Moderate response rate											
		Pattern 1				Pattern 2				Pattern 3			
Metric	Scenario	IPSW	PSS	MI5	MI100	IPSW	PSS	MI5	MI100	IPSW	PSS	MI5	MI100
Bias ***x10^3***													
	1.1	-3	-2	3	-2	31	37	19	6	553	750	101	37
	1.2	1	1	3	1	53	53	65	54	1554	1549	1611	1549
	2.1	-4	-3	4	-2	-9	-7	10	-4	180	346	58	-11
	2.2	13	5	17	12	169	32	-41	-52	-551	359	-95	-150
RMSE ***x10^3***													
	1.1	91	92	100	91	140	125	113	108	1008	1028	604	583
	1.2	90	92	104	91	106	106	114	106	1631	1628	1691	1627
	2.1	93	94	98	94	141	126	138	130	909	839	726	687
	2.2	94	94	106	94	286	132	115	118	1363	808	595	601
95% CI cov.													
	1.1	95	95	96	95	93	92	93	96	73	71	93	96
	1.2	95	95	96	95	91	92	88	91	12	12	12	11
	2.1	96	96	95	96	94	95	93	96	90	88	94	95
	2.2	96	96	96	96	76	92	90	92	76	87	93	94

Table3. Performance Metrics by Pattern, Levels of Associations with Auxiliary Variables (500 simulated data sets of n = 3,000)

		Low response rate											
		Pattern 1				Pattern 2				Pattern 3			
Metric	Scenario	IPSW	PSS	MI5	MI100	IPSW	PSS	MI5	MI100	IPSW	PSS	MI5	MI100
Bias ***x10^3***													
	1.1	2	3	1	2	26	45	14	2	577	970	70	12
	1.2	9	10	7	9	101	102	102	101	3001	3002	2994	2992
	2.1	3	8	8	2	7	15	22	4	172	499	112	21
	2.2	35	16	34	34	1028	75	-134	-145	-28	770	-565	-620
RMSE ***x10^3***													
	1.1	32	32	29	32	157	123	63	57	1113	1187	322	289
	1.2	33	34	30	32	107	108	106	107	3007	3008	2998	2998
	2.1	33	34	34	32	148	119	93	75	939	864	470	380
	2.2	48	37	45	47	1074	123	145	154	1607	922	636	677
95% CI cov.													
	1.1	94	94	88	93	83	85	88	94	57	53	88	94
	1.2	93	92	88	91	16	17	14	13	0	0	0	0
	2.1	95	95	92	94	89	93	92	94	76	77	92	94
	2.2	82	92	72	81	25	85	34	23	39	59	51	38

Continued

Table 3. Continued

		Moderate response rate											
		Pattern 1				Pattern 2				Pattern 3			
Metric	Scenario	IPSW	PSS	MI5	MI100	IPSW	PSS	MI5	MI100	IPSW	PSS	MI5	MI100
Bias $x10^3$													
	1.1	0	1	0	0	21	11	3	0	221	192	14	3
	1.2	4	4	2	4	52	53	51	51	1537	1540	1523	1523
	2.1	0	2	1	0	-2	-4	4	0	32	74	23	2
	2.2	17	7	16	16	574	41	-44	-46	-1001	323	-117	-124
RMSE $x10^3$													
	1.1	23	23	24	24	80	76	29	28	552	529	153	150
	1.2	24	24	24	24	57	58	56	56	1543	1545	1528	1529
	2.1	24	24	26	24	49	43	33	31	343	275	174	166
	2.2	29	25	29	29	636	52	52	53	1548	370	192	195
95% CI cov.													
	1.1	94	94	93	94	92	94	93	95	69	80	94	95
	1.2	94	94	95	94	40	36	39	40	0	0	0	0
	2.1	94	94	94	94	95	94	91	94	91	94	92	95
	2.2	88	93	88	88	37	74	65	64	70	55	86	87

Table 4. Performance Metrics by Pattern, Levels of Associations with Auxiliary Variables (500 simulated data sets of n = 10,000)

		Low response rate											
		Pattern 1				Pattern 2				Pattern 3			
Metric	Scenario	IPSW	PSS	MI5	MI100	IPSW	PSS	MI5	MI100	IPSW	PSS	MI5	MI100
Bias ***x10^3***													
	1.1	1	1	-1	0	20	37	6	1	467	730	32	4
	1.2	7	7	5	7	101	101	99	99	3000	3002	2984	2980
	2.1	1	5	3	0	3	1	10	2	128	208	53	11
	2.2	33	12	31	32	1349	74	-144	-147	181	755	-614	-632
RMSE ***x10^3***													
	1.1	18	18	17	18	122	102	39	31	895	928	196	159
	1.2	19	19	16	19	103	103	101	101	3002	3004	2985	2982
	2.1	18	19	22	18	96	89	52	40	615	604	261	204
	2.2	37	22	36	37	1384	89	148	150	1584	799	639	650
95% CI cov.													
	1.1	97	97	94	96	86	85	94	96	56	53	95	95
	1.2	93	93	87	93	0	0	0	0	0	0	0	0
	2.1	97	95	96	96	94	96	96	94	82	84	97	94
	2.2	55	90	55	55	14	66	1	0	48	18	7	3

Continued

Table 4. Continued

		Moderate response rate											
		Pattern 1				Pattern 2				Pattern 3			
Metric	Scenario	IPSW	PSS	MI5	MI100	IPSW	PSS	MI5	MI100	IPSW	PSS	MI5	MI100
Bias $x10^3$													
	1.1	1	1	0	0	0	6	2	0	53	132	9	2
	1.2	4	5	3	4	51	52	50	51	1533	1536	1516	1521
	2.1	1	3	2	0	-1	-5	2	0	3	69	10	0
	2.2	17	6	17	17	722	41	-45	-46	-1229	326	-123	-127
RMSE $x10^3$													
	1.1	13	13	13	13	66	44	17	15	514	320	90	81
	1.2	14	14	13	13	53	54	51	52	1535	1537	1518	1522
	2.1	13	13	15	13	36	23	20	17	251	157	105	90
	2.2	22	15	21	21	773	45	48	48	1694	340	151	151
95% CI cov.													
	1.1	96	95	95	95	91	92	96	96	70	89	96	95
	1.2	94	94	94	94	3	2	4	3	0	0	0	0
	2.1	95	94	97	95	95	94	97	95	94	95	97	95
	2.2	75	92	75	76	14	33	19	14	61	8	70	65

PATTERN 1: AUXILIARY VARIABLES WEAKLY ASSOCIATED WITH R AND Y

Tables 2-4 show bias, RMSE, and 95 percent CI coverage of the estimated mean outcome Y for different scenarios using PSW and MI methods. As expected, when auxiliary variables are weakly associated with R and Y, the nonresponse adjustment is weak, and therefore the choice of adjustment method matters only for scenario 2.2. Even when adjustment is not strong, PSS outperforms MI in scenario 2.2 as the sample size gets larger; for sample size 10,000, the RMSE of PSS is about 30 percent lower than that of MI, and its coverage is 90 percent compared to 55 percent for MI under a low response rate (see the upper left quadrant of table 4).

PATTERN 2: AUXILIARY VARIABLES STRONGLY ASSOCIATED WITH R, BUT WEAKLY ASSOCIATED WITH Y

The second pattern addresses a common situation among survey researchers: including variables that are strongly associated with the response, but not with the survey variables of interest in weighting adjustment, which can unduly increase the variance in weighting adjustments. We hypothesized that global nonresponse adjustment models in social surveys with multiple outcomes are likely to include auxiliary variables that are not associated with all Y variables; therefore, PSW may be unduly increasing variance in certain Y variables. Results from the simulations show that MI tends to result in smaller penalties on the variance (not shown in the tables) in pattern 2 when the adjustment model is correctly specified (scenarios 1.1 and 2.1) and the sample size is large (n = 3,000 or 10,000). However, looking at the coverage of confidence intervals, we see that MI performs only slightly better than PSW in the correctly specified main effects model (scenario 1.1), and it generally produces similar coverage rates for the correctly specified complex model (scenario 2.1).

When adjustment models are complex and misspecified, it becomes harder to draw conclusions about the methods. It is clear that IPSW in the misspecified complex model (scenario 2.2) performs poorly, producing standard errors of up to ten times larger than those of the unadjusted estimates (compared to only 1.17 for MI with five imputations) and leading to lower bias reduction compared to MI methods for n = 3,000. That said, the other weighting method, PSS, may produce lower RMSE and better coverage in the misspecified complex model. For example, PSS has better coverage of confidence intervals compared to MI methods in both the low response (PSS = 85 percent compared to MI5 = 34 percent and MI100 = 23 percent) and the moderate response cases (PSS = 74 percent compared to MI5 = 65 percent and MI100 = 64 percent). However, comparing the corresponding results from the small sample size (n = 200), we see that MI produces slightly higher variance estimates and similar coverage rates compared to PSS.

PATTERN 3: AUXILIARY VARIABLES STRONGLY ASSOCIATED WITH BOTH R AND Y

Having auxiliary variables that are strongly associated with both R and Y provides the ideal case for effective adjustment. In this case, there would be substantial bias in the unadjusted mean estimates, and adjustment by MI and PSW would be strong. The third pattern illustrates this particular case. The RMSE for this pattern favors MI over PSW except for scenario 1.2, where they perform similarly poorly, and except for the small sample size (n = 200), where results are rather mixed. We also find that MI maintains better coverage compared to PSS in pattern 3 under low response rates when the adjustment models are correctly specified. However, this advantage disappears or is no longer consistent when adjustment models are misspecified.

Furthermore, in this pattern, misspecification of the adjustment model by omitting interactions and quadratic terms (scenario 2.2) may have a dramatic effect on the variance estimates from IPSW in this pattern. As Kang and Schafer (2007, 529) described for item nonresponse, our results suggests that "in practice, a good data analyst would never use a simple inverse propensity score weighted estimator if the weights were too extreme. Unusually large weights may be taken as a sign of model failure, prompting the researcher to revise the n (propensity score) model." As such, the misspecified complex model produced weights of up to 900.

Figure 2 provides a closer look at variance inflation by PSW and MI methods, and shows the relative standard errors of the adjusted respondent mean for different scenarios and levels of associations (patterns) under a low response rate and n = 3,000. Each symbol represents the ratio of the average standard errors for a given method divided by the standard error of the unadjusted respondent mean. A value of 1 implies no change in standard error after applying nonresponse adjustment, while values above or below 1, respectively, indicate an increase or decrease in standard error relative to that of the unadjusted mean estimates. For the correctly specified models (scenarios 1.1 and 2.1), variance inflation from adjustment with MI remains lower than that from adjustment with PSW methods. In the first case of misspecification where the main-effects model is misspecified by omitting one auxiliary variable (scenario 1.2), all adjustment methods perform similarly, and variance inflation remains low. Conversely, misspecification of the complex true response and outcome models by omitting complex terms results in extreme IPSW weights and substantial inflation in standard errors. To put variance inflation and bias into perspective, it is necessary to also look at the CI coverage of the methods. Figure 3 identifies the best MI and PSW methods with respect to CI coverage, comparing PSS and MI5 for the same sample size of 3,000, as shown in figure 2. Two findings are noteworthy. First, regardless of the method, response rate, and sample size, omitting an important auxiliary variable (highly associated with both R and Y, scenario 1.2 under pattern 3) results in 0 coverage.

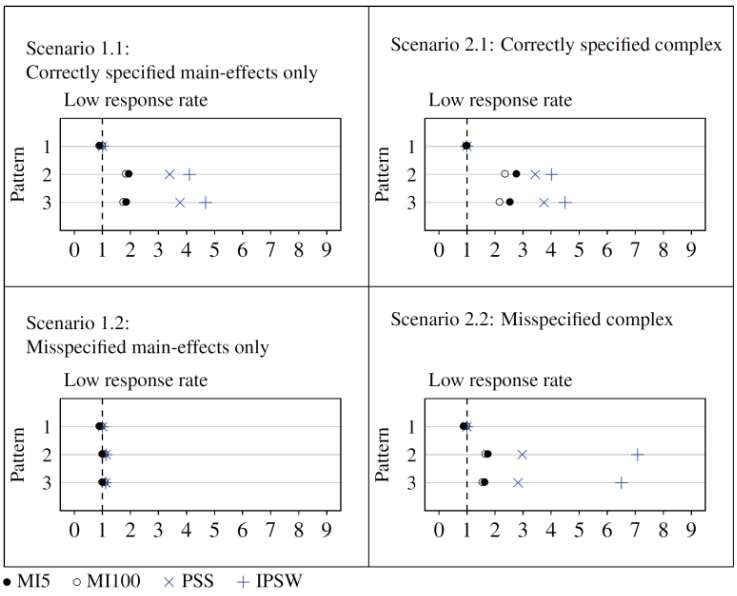

Figure 2. Relative Ratio of Standard Error for the Adjusted Respondent Mean to Unadjusted Respondent Mean, n = 3,000.

This is due to high bias in the mean estimates caused by the omission of an auxiliary variable strongly associated with both R and Y. If the omitted variable is strongly related to R but not Y (scenario 1.2 under pattern 2), coverage still suffers, but it improves with higher response rate or with smaller sample size (results for other sample sizes are not shown in the figure). Second, PSS produces better coverage than MI in scenario 2.2 under a low response rate, although the results are mixed under a moderate response rate. Considered together with figure 2, figure 3 indicates that while MI provides smaller variance estimates, it may result in lower coverage. Moreover, the trade-offs between variance, bias, and coverage vary substantially by sample size and response rate (see tables 2-4 and supplementary data online).

IPSW VERSUS PSS

IPSW has lower bias compared to PSS under certain conditions (e.g., scenarios 1.1 and 2.1). However, the variance for IPSW is higher, meaning it is better on RMSE only under specific cases. Furthermore, IPSW almost never has higher coverage than PSS, with the exception of pattern 3 in scenario 1.1, where IPSW has higher coverage for a low response rate or low sample size (n = 200). Also, as the sample size increases with a moderate response rate, PSS fares worse and

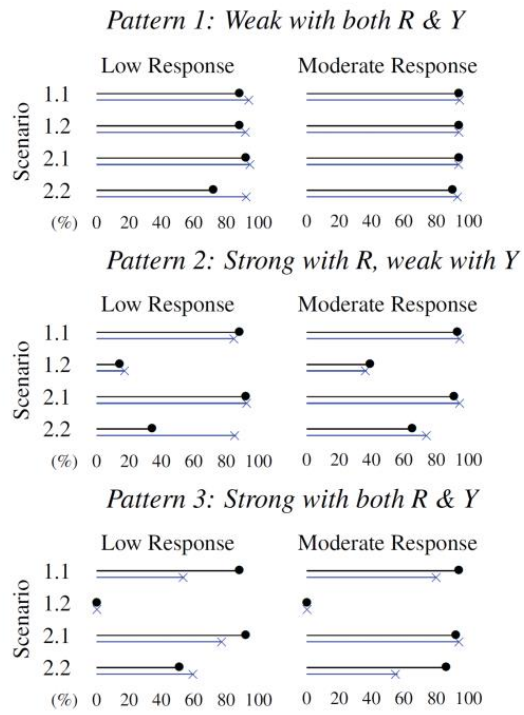

Figure 3. 95% CI Coverage of PSS (-x) and MI5 (-•), n = 3,000.

worse in pattern 3 in scenario 2.2, allowing IPSW to have higher coverage for n = 10,000 and under moderate response rates for n = 3,000.

MI5 VERSUS MI100

RMSE for MI100 are generally lower than for MI5 when adjustment models are correctly specified (e.g., scenarios 1.1 and 2.1), but the gains in efficiency tend to be more modest under moderate response rates or larger sample sizes. When adjustment models are complex and misspecified, however, MI5 can sometimes be more advantageous than MI100 (e.g., lower RMSE under scenario 2.2 and low response rate for n = 3,000 and 10,000). With regard to coverage, MI5 performs better under scenario 2.2, patterns 2 and 3, when the response rate is low.

Overall, the most interesting comparisons are between PSS and MI5. While it is unlikely, if pattern 3 is achieved, MI5 is preferred on RMSE and coverage. None of the methods perform well under scenario 1.2, where an auxiliary variable is omitted. For patterns 1 and 2, MI5 is preferred under scenario 1.1, while PSS is preferred under scenario 2.2, whereas neither is clearly better under scenario 2.1. In conclusion, neither of the PSW and MI methods can be consistently preferred for unit-nonresponse adjustment.

Conclusion and Discussion

The simulation results highlight the advantages of PSW and MI under different scenarios. However, as neither of these consistently outperforms the other method, we advocate the use of robust methods that simultaneously model both the outcomes and the (non)response. However, we recommend that researchers avoid IPSW, as its performance is generally poor in the misspecified complex model. On the other hand, while overall MI5 and MI100 perform similarly, MI100 has lower coverage in some cases, specifically when the response rate is low and adjustment models are complex but misspecified in patterns 2 and 3. This is sometimes caused by the lower variance estimates, and other times both higher bias and lower variance estimates for MI100 compared to MI5.

Although MI performs better than or similar to PSW in relatively more instances, sub-classification may also provide lower bias and better coverage. For example, when true models are complex and misspecified, and auxiliary variables are strongly related to R but not Y, MI yields lower standard errors and higher bias compared to PSS, resulting in lower coverage of the 95 percent confidence intervals under n = 3,000 or 10,000. Therefore, we expect methods that combine propensity scores with MI (e.g., Jolani, van Buuren, and Frank 2011) or those modeling response and outcome simultaneously (Little and An 2004; Zhang and Little 2011) to be attractive alternatives in the future as they become more commonly available to practitioners. Jolani and his colleagues show that combined propensity score and MI methods perform well under correct model specification, but more exploration is needed for misspecified models. Similarly, a number of studies explore robust methods, modeling response and outcome simultaneously under correct and misspecified models. For example, Kang and Schafer (2007, 532) showed that “two wrong [misspecified] models are not necessarily better than one,” referring to models for the propensity score and the regression of outcome on auxiliary variables. Some other studies offered modifications to doubly robust methods, which seem to yield more favorable results even if neither model is correctly specified (e.g., Cao, Tsiatis, and Davidian 2009). Therefore, we are looking forward to the incorporation of these combined applications or dual modeling approaches to the mainstream toolbox of practitioners.

We evaluated the relative performance of unit-nonresponse adjustment with PSW and MI methods. The results show that MI can be a working alternative to PSW in unit-nonresponse adjustments with its strengths and weaknesses, as illustrated in the simulation results. However, it is worth noting that MI requires meticulous modeling, and its application may raise several user-oriented practical issues. First of all, model misspecification, such as “uncongenial imputation,” where the imputation model is not as rich as the analytical model, can cause biased estimates in multivariate analysis (Meng 1994). Therefore, employing MI requires more caution especially while generating global nonresponse weights for social surveys. Another concern with MI is that it may require more effort and expertise in model specification. Social surveys usually include categorical, ordinal, and/or count variables, which do not meet the assumption of multivariate normality required in standard multiple-imputation procedures. In addition, for all scenarios discussed here, we assumed MAR. However, there are also methods that allow the missingness to be not at random (MNAR) as proposed by, for example, Diggle and Kenward (1994), and Molenberghs, Kenward, and Lesaffre (1997). These methods have not yet become common practice and are still open to debate.

Second, even though MI could be a viable alternative to PSW, the user-oriented practical issues such as expertise, accessibility of auxiliary data, and timeliness remain essential for the future use of MI unit-nonresponse adjustments. A major problem with MI is to ensure that the imputation model is congenial to the analysis undertaken. So, if the information and resources available to the analysts and the imputer are different, imputers' input is needed at the analysis stage. As social surveys often involve a number of outcome variables, it may not be possible to consider all interactions, all subgroups, and any other aspects of interest to analysts. Then, the imputer needs to be available to advise analysts or make the limitations of the imputation model clear to the users. The other option is to provide the auxiliary data available for weighting so that analysts can build their own imputation models. Either way, MI seems to cost extra time since it requires the modeling of all variables related to the analytical model.

Another practical concern could be the additional adjustment required for the application of MI in complex surveys. It may not be straightforward to determine how MI can be combined with complex design weights (van Buuren 2012) if design variables are not provided in the public data. However, there have been improvements in this field. Alternative methods are now available, including design variables (or weights in the worst case) in the specification of imputation models (e.g., Reiter, Raghunathan, and Kinney 2006) and more complex methods offering new modeling approaches and combining rules (Zhou, Raghunathan, and Elliott 2012). These methods, however, are designed for item nonresponse, and further exploration is needed for their applications in unit nonresponse. There are certainly other practical concerns regarding the

use of MI in unit-nonresponse adjustments, but these are beyond the scope of this study. More advice on the modeling aspect of applications is provided by Little (1988).
Regarding IPSW, some suggest that extreme weights result from logistic regression rather than being a problem with the method itself (Ridgeway and McCaffrey 2007). Therefore, the use of alternative methods such as generalized boosted models described by McCaffrey, Ridgeway, and Morral (2004) for propensity-score estimation may prevent probabilities close to 0 or 1, and yield lower variance and bias estimates for IPSW. This can be a topic for further research. However, doubly robust methods are still not commonplace in survey methodology, and in many applications of IPSW, weights are obtained by logistic regression.
Our simulations compared MI and PSW for mean estimates of continuous outcomes. This study could be extended to handle binary and ordinal variables, and to evaluate other parameters such as regression coefficients or domain means. Also, future research could examine the relative performance of MI to PSW in the joint estimation of means of several outcomes. In the multivariate case with several outcomes, MI allows for modeling auxiliary-outcome relationships correctly. This can lead to an extra efficiency of MI in unit-nonresponse adjustment, which comes from the fact that different outcomes are related differently to the weights. For example, one Y variable in a social survey might be in scenario 1.1 of our simulations and another could be in scenario 2.1. In this case, to adjust by weighting, the same model would be fitted for all outcomes, although their respective relationships with unit nonresponse vary, resulting in inefficient estimates. This, however, is a mixed blessing, since MI requires careful modeling of all the variables. Overall, we do not definitively conclude which method offers a better bias-variance trade-off. More exploration is needed with complex models and real survey-data applications and doubly robust methods before giving concrete recommendations.

Appendix 1. Coefficients for Data-Generation Models by Patterns

1. WEAK WITH BOTH R AND Y

$$\alpha_0 = -0.9 \text{ for low}, 0.5 \text{ for moderate response rate}$$

$$\alpha_1 = 0.1, \alpha_2 = 0.1\ (\alpha_3 = 0.1, \alpha_4 = 0.1, \alpha_5 = 0.1)$$

$$\beta_1 = 0.1, \beta_2 = 0.1, \beta_e = 1\ (\beta_3 = 0.1, \beta_4 = 0.1, \beta_5 = 0.1)$$

2. STRONG WITH R AND WEAK WITH Y

$$\alpha_0 = -1.8 \; for \quad or2 \; for \; moderate \; response \; rate$$

$$\alpha_1 = 2, \alpha_2 = 4 \; (\alpha_3 = 0.5, \alpha_4 = 1.5, \alpha_5 = 0.8)$$

$$\beta_1 = 0.1, \beta_2 = 0.1, \beta_e = 1 \; (\beta_3 = 0.1, \beta_4 = 0.1, \beta_5 = 0.1)$$

3. STRONG WITH BOTH R AND Y

$$\alpha_0 = -1.8 \; for \quad or2 \; for \; moderate \; response \; rate$$

$$\alpha_1 = 2, \alpha_2 = 4 \; (\alpha_3 = 0.5, \alpha_4 = 1.5, \alpha_5 = 0.8)$$

$$\beta_1 = 1, \beta_2 = 3, \beta_e = 5 \; (\beta_3 = 0.3, \beta_4 = 0.8, \beta_5 = 0.5)$$

Appendix 2. Resulting Mean Response Rates from Data-Generation Models (%)

Scenario	Response rate	Pattern 1	Pattern 2	Pattern 3
1.1, 1.2				
	Low	33	35	35
	Moderate	62	66	66
2.1, 2.2				
	Low	33	37	37
	Moderate	62	65	65

Supplementary Data

Supplementary data are freely available online at http://poq.oxfordjournals. org/.

References

Bethlehem, Jelke, Fannie Cobben, and Barry Schouten. 2011. Handbook of Nonresponse in Household Surveys. Hoboken, NJ: Wiley.

Beunckens, Caroline, Cristina Sotto, and Geert Molenberghs. 2008. “A Simulation Study Comparing Weighted Estimating Equations with Multiple Imputation Based Estimating Equations for Longitudinal Binary Data.” Computational Statistics & Data Analysis 52:1533-1548.

Brookhart, M. Alan, Sebastian Schneeweiss, Kenneth J. Rothman, Robert J. Glynn, Jerry Avorn, and Til Stürmer. 2006. “Variable Selection for Propensity Score Models.” American Journal of Epidemiology 163:1149-1156.
Cao, Weihua, Anastasios A. Tsiatis, and Marie Davidian. 2009. “Improving Efficiency and Robustness of the Doubly Robust Estimator for a Population Mean with Incomplete Data.” Biometrika 96:723-34.
Carpenter, James R., Michael G. Kenward, and Stijn Vansteelandt. 2006. “A Comparison of Multiple Imputation and Doubly Robust Estimation for Analyses with Missing Data.” Journal of the Royal Statistical Society: Series A (Statistics in Society) 169:571-84.
Clarke, Kevin A., Brenton Kenkel, and Miguel R. Rueda. 2011. “Misspecification and the Propensity Score: the Possibility of Overadjustment.” Unpublished manuscript.
D'Agostino, Ralph B. Jr., and Donald B. Rubin. 2000. “Estimating and Using Propensity Scores with Partially Missing Data.” Journal of the American Statistical Association 95:749-59.
Diggle, Peter J., and Michael G. Kenward. 1994. “Informative Dropout in Longitudinal Data Analysis (with discussion).” Journal of the Royal Statistical Society: Series C 43:49-93.
Drake, Christiana. 1993. “Effects of Misspecification on Propensity Score Estimators of Treatment Effect.” Biometrics 49:1231-1236.
Groves, Robert M. 2006. “Nonresponse Rates and Nonresponse Bias in Household Surveys.” Public Opinion Quarterly 70:646-75.
Hirano, Keisuke, and Guido W. Imbens. 2001. “Estimation of Causal Effects Using Propensity Score Weighting: An Application to Data on Right Heart Catheterization.” Health Services and Outcomes Research Methodology 2:259-78.
Jolani, Shahab, Stef van Buuren, and Laurence E. Frank. 2011. “Combining the Complete-Data and Nonresponse Models for Drawing Imputations under MAR.” Journal of Statistical Computation and Simulation 83(5):868-879.
Kang, Joseph D. Y., and Joseph L. Schafer. 2007. “Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data.” Statistical Science 22:523-39.
Kreuter, Frauke, and Kristen Olson. 2011. “Multiple Auxiliary Variables in Nonresponse Adjustment.” Sociological Methods & Research 40:311-32.
Kreuter, Frauke, Kristen Olson, James Wagner, Ting Yan, Trena M. EzzatiRice, Carolina CasasCordero, Michael Lemay, Andy Peytchev, Robert M. Groves, and Trivellore E. Raghunathan. 2010. “Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-Response: Examples from Multiple Surveys.” Journal of the Royal Statistical Society: Series A (Statistics in Society) 173:389-407.
Lee, Brian K., Justin Lessler, and Elizabeth A. Stuart. 2011. “Weight Trimming and Propensity Score Weighting.” PloS One 6.3:e18174.
Lee, Katherine J., and John B. Carlin. 2010. “Multiple Imputation for Missing Data: Fully Conditional Specification versus Multivariate Normal Imputation.” American Journal of Epidemiology 171:624-32.
Lee, Sunghee, and Richard Valliant. 2009. “Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment.” Sociological Methods & Research 37:319-43.
Little, Roderick J. A. 1986. “Survey Nonresponse Adjustments.” International Statistical Review 54:3.
———. 1988. “Missing-Data Adjustments in Large Surveys.” Journal of Business & Economic Statistics 6:287-96.
———. 2013. Discussion. Journal of Official Statistics 29:363-66.
Little, Roderick J. A., and Hyonggin An. 2004. “Robust Likelihood-Based Analysis of Multivariate Data with Missing Values.” Statistica Sinica 14:949-68.

Little, Roderick J. A., and Donald B. Rubin. 2002. Statistical Analysis with Missing Data. New York: John Wiley & Sons.
Little, Roderick J. A., and Sonya Vartivarian. 2005. “Does Weighting for Nonresponse Increase the Variance of Survey Means?” Survey Methodology 31:161-68.
Matsuo, Hideko, Jaak Billiet, Geert Loosveldt, Frode Berglund, and 0yven Kleven. 2010. “Measurement and Adjustment of Non-Response Bias Based on Non-Response Surveys: The Case of Belgium and Norway in the European Social Survey Round 3.” Survey Research Methods 4:165-78.
Mattei, Alessandra. 2009. “Estimating and Using Propensity Score in Presence of Missing Background Data: an Application to Assess the Impact of Childbearing on Wellbeing.” Statistical Methods and Applications 18:257-73.
McCaffrey, Daniel F., Greg Ridgeway, and Andrew R. Morral. 2004. “Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies.” Psychological Methods 9:403.
Meng, Xiao-Li. 1994. “Multiple-Imputation Inferences with Uncongenial Sources of Input.” Statistical Science 9:538-58.
Millimet, Daniel L., and Rusty Tchernis. 2009. “On the Specification of Propensity Scores, with Applications to the Analysis of Trade Policies.” Journal of Business & Economic Statistic 27:397-415.
Molenberghs, Geert, Michael G. Kenward, and Emmanuel Lesaffre. 1997. “The Analysis of Longitudinal Ordinal Data with Non-Random Dropout.” Biometrika 84:33-44.
Peytchev, Andy. 2012. “Multiple Imputation for Unit Nonresponse and Measurement Error.” Public Opinion Quarterly 76:214-37.
Qu, Yongming, and Ilya Lipkovich. 2009. “Propensity Score Estimation with Missing Values Using a Multiple Imputation Missingness Pattern (MIMP) Approach.” Statistics in Medicine 28:1402-1414.
Raghunathan, Trivellore E., James M. Lepkowski, John Van Hoewyk, and Peter Solenberger. 2001. “A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models.” Survey Methodology 27:85-96.
Reiter, Jerome P., Trivellore E. Raghunathan, and Satkartar K. Kinney. 2006. “The Importance of Modeling the Sampling Design in Multiple Imputation for Missing Data.” Survey Methodology 32:143-49.
Ridgeway, Greg, and Daniel F. McCaffrey. 2007. “Comment: Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data.” Statistical Science 22:540-43.
Rosenbaum, Paul R., and Donald B. Rubin. 1984. “Reducing Bias in Observational Studies Using Subclassification on the Propensity Score.” Journal of the American Statistical Association 79:516-24.
Rubin, Donald B. 1976. “Inference and Missing Data.” Biometrika 63:581-92.
———. 1987. Multiple Imputation for Nonresponse in Surveys. New York: Wiley.
Schafer, Joseph L. 1999. “Multiple Imputation: A Primer.” Statistical Methods in Medical Research 8:3-15.
Setoguchi, Soko, Sebastian Schneeweiss, M. Alan Brookhart, Robert J. Glynn, and E. Francis Cook. 2008. “Evaluating Uses of Data Mining Techniques in Propensity Score Estimation: A Simulation Study.” Pharmacoepidemiology and Drug Safety 17:546-55.
Stoop, Ineke, Jaak Billiet, Achim Koch, and Rory Fitzgerald. 2010. Improving Survey Response: Lessons Learned from the European Social Survey. New York: Wiley.
Van Buuren, Stef. 2012. Flexible Imputation of Missing Data. Boca Raton, FL: CRC Press. Wagner, James. 2012. “A Comparison of Alternative Indicators for the Risk of Nonresponse Bias.”
Public Opinion Quarterly 76:555-75.
White, Ian R., Patrick Royston, and Angela M. Wood. 2011. “Multiple Imputation Using Chained Equations: Issues and Guidance for Practice.” Statistics in Medicine 30:377-99.

Yuan, Ying, and Roderick J. A. Little. 2007. “Model-Based Estimates of the Finite Population Mean for Two-Stage Cluster Samples with Unit Non-Response.” Journal of the Royal Statistical Society: Series C (Applied Statistics) 56:79-97.
Zhang, Guangyu, and Roderick J. A. Little. 2011. “A Comparative Study of Doubly Robust Estimators of the Mean with Missing Data.” Journal of Statistical Computation and Simulation 81:2039-2058.
Zhou, Hanzhi, Trivellore E. Raghunathan, and Michael R. Elliott. 2012. “A Semi-Parametric Approach to Account for Complex Designs in Multiple Imputation.” https://fcsm.sites.usa.gov/files/2014/05/Zhou_2012FCSM_X-A.pdf.